DOCUMENT RESUME

ED 457 192 TM 033 289

AUTHOR Lambert, Richard G.; Flowers, Claudia
TITLE A Procedure for Testing the Difference between Effect Sizes.
PUB DATE 1998-04-00
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Effect Size; Monte Carlo Methods; *Pretests Posttests; Simulation
IDENTIFIERS Type I Errors

ABSTRACT

A special case of the homogeneity of effect size test, as applied to pairwise comparisons of standardized mean differences, was evaluated. Procedures for comparing pairs of pretest to posttest effect sizes, as well as pairs of treatment versus control group effect sizes, were examined. Monte Carlo simulation was used to generate Type I error rates and power values for tests of the differences in independent effect sizes based on both the "g" and "d" methods. Type I error rate was evaluated by crossing 6 sample size conditions (5, 10, 20, 30, 50, and 100) by 5 population effect size conditions (0.00, 0.25, 0.50, 0.75, and 1.00). Power was evaluated by crossing the 6 sample size conditions by 4 conditions representing the magnitude of the difference between treatment and control conditions (0.25, 0.50, 0.75, and 1.00). The "d" based statistic yielded Type I error rates closer to the nominal level than did the "g" based statistic while yielding a slightly more conservative method for testing the difference between two effect size measures. Examples are provided that illustrate the use of these procedures as post hoc comparison techniques following factorial analysis of variance designs. (Contains 6 tables and 30 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 457 192

A Procedure for Testing the Difference between Effect Sizes

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

R. Lambert

Richard G. Lambert

Claudia Flowers

University of North Carolina at Charlotte

Paper presented at the Annual Meetings of the American Educational Research Association, April, 1998, San Diego, CA.

Correspondence concerning this article should be addressed to Richard G. Lambert, University of North Carolina at Charlotte, Department of Educational Administration, Research, and Technology, 3135 Colvard, Charlotte, NC 28223-0001. Electronic correspondence may be sent to rglamber@email.uncc.edu.

TM033289

BEST COPY AVAILABLE

Author Footnotes

Correspondence concerning this article should be addressed to Richard G. Lambert, University of North Carolina at Charlotte, Department of Educational Administration, Research, and Technology, 3135 Colvard, Charlotte, NC 28223-0001. Electronic correspondence may be sent to rglamber@email.uncc.edu.

Abstract

A special case of the homogeneity of effect size test, as applied to pairwise comparisons of standardized mean difference effect sizes, was evaluated. Procedures for comparing pairs of pretest to posttest effect sizes, as well as pairs of treatment versus control group effect sizes were examined. Monte Carlo simulation was used to generate Type I error rates and power values for tests of the differences in independent effect sizes based on both the g and d methods. Type I error rate was evaluated by crossing six sample size conditions (5, 10, 20, 30, 50, and 100) by five population effect size, δ, conditions (.00, .25, .50, .75, and 1.00). Power was evaluated by crossing the six sample size conditions by four conditions representing the magnitude of the difference between treatment, δ_1, and control, δ_2, conditions (.25, .50, .75, and 1.00). The d based statistic yielded Type I error rates closer to the nominal level than did the g based statistic while yielding a slightly conservative method for testing the difference between two effect size measures. Examples are provided which illustrate the use of these procedures as posthoc comparison techniques following factorial ANOVA designs.

Testing the Difference between Effect Sizes as a Posthoc Comparison Procedure Following Factorial ANOVA

There has been a well-documented shift in emphasis within the educational research literature away from reporting only traditional statistical significance testing toward reporting measures of effect magnitude. The fourth edition of the APA publication guidelines (1994) suggest that authors of primary studies include in their dissemination efforts either effect sizes or information sufficient to reconstruct them. The APA Task Force on Statistical Inference (Wilkinson & The APA Task Force on Statistical Inference, 1999) strongly urges researchers to supplement the reporting of p values with effect size information. The fifth edition of the APA publication guidelines (2001) make an even stronger statement, declaring that it is always necessary to include effect size measures when reporting the results of a quantitative study. Furthermore, thorough reporting of the results of quantitative primary studies includes the calculation and interpretation of effect size information, both to facilitate meta-analytic synthesis and to describe the findings in a complete and accessible format (Thompson, 1996). At the same time, the use of traditional null hypothesis significance testing has been widely questioned and criticized (Thompson, 1993; Falk & Greenbaum, 1995; Kirk, 1996; Harlow, Muliak, and Stieger, 1997).

Tukey (1969) contrasted the "sanctification" process of significance testing with the real "detective work" of scientific inquiry. Similarly, Fan (2001) has argued that statistical significance testing has been given an artificially high and somewhat misguided position of reverence and sanctity by educational researchers. In addition, he argues that many educational researchers falsely believe they are exempt from considerations of sampling variation when

reporting effect sizes. He further demonstrates that both statistical significance tests and effect sizes are needed to fully interpret an educational experiment, are related measures serving different purposes, and complement rather than substitute for each other. As Levin (1993) has argued, statistical significance testing is still needed to enable educational researchers to correctly interpret their results, including their effect sizes.

Furthermore, within the field of meta-analysis itself, suggestions have been put forth regarding the need to move beyond descriptive meta-analyses toward syntheses which involve hypothesis testing and theory development (Becker, 1989; Becker & Scram, 1994; Miller & Pollock, 1994). When conducting hypotheses driven meta-analyses, researchers can be faced with the need to test differences between summarized effect size estimates (Alliger, 1995). Similarly, when estimates are obtained from theoretically meaningful subsets of effect sizes within descriptive meta-analyses, the condition of heterogeneity of population effect sizes can be tested (Hedges & Olkin, 1985) and the resulting differences can represent valuable information to practitioners. All of these situations involve the use of the sampling distributions of effect sizes and significance tests based on their properties.

As journal editors begin to require greater use of measures of effect magnitude and educational researchers become more familiar with the use and interpretation of effect sizes, opportunities to place confidence intervals around effect sizes and to directly test the difference between effect sizes will present themselves. The purpose of this paper is to evaluate a procedure for testing the difference between pairs of effect sizes in the context of posthoc comparisons following factorial ANOVA.

The need to examine the difference between pairs of effect sizes within a single study

may present itself when an educational researcher has used a factorial ANOVA design and has found a statistically significant interaction effect. While the interpretation of interactions continues to be a source of difficulty for many students of educational research (Oshima & McCarty, 1999), several useful analogies are available. For example, a researcher can observe whether parallel lines appear when line graphs of cell means are created. Each line may represent all the cell means within a given level of one factor. When the lines created for all the levels of that factor are compared, parallel lines indicate no interaction while non-parallel lines represent interaction. Another way to think about the concept of interaction is in terms of the differences between differences. When the differences between the cell means within one level of a factor are greater than the differences between cell means within at least one other level of the same factor, then interaction may be present. Since the differences between cell means within the levels of a given factor may be expressed in effect sizes, then a test for the differences between such effect sizes may be useful in interpreting interaction effects.

Educational researchers are often interested in the interaction effects in ANOVA designs. The researcher's primary focus in studies with pretests and posttests is often to determine whether the effect magnitude across time in the treatment group exceeds that of the control group. Similarly, a researcher's primary focus in a study with only between-subjects terms is often whether the effect magnitude between a particular pair of cells exceeds that of another pair of cells. However, the use of simple effects and traditional post hoc comparison techniques does not always offer a direct answer to this research question regarding group differences in effect magnitude. Several post hoc tests are often needed to address what may be the central focus of the study. This article will attempt to evaluate a single effect size based significance test that may

be used to directly compare the effect size in the treatment group to that of the control or comparison group. In addition, researchers may at times be interested only in the interaction effects in a factorial ANOVA design. In such cases they may elect to proceed directly to a procedure such as the one described here.

The Independent Group Means Case

The independent case involves testing the difference between independent effect sizes, each calculated from two independent group means. This method assumes that the means utilized in calculating the individual effect sizes are independent and that the effect sizes being compared are independent. Factorial designs involving between subjects variables are commonly used within educational research. The procedure described here will compare the difference between pairs of treatment versus control group standardized mean difference effect sizes. The need for this procedure arises when, for example, an educational researcher is interested in investigating whether the effect size for some achievement variable, when comparing the treatment and control groups, is different for male and female students. Similarly, treatment versus control effect sizes might be calculated and then compared across the levels of a variety of stratification or blocking variables such as high versus low ability on another measure of interest. If statistically significant interactions are found in such designs, an effect size based significance test could allow the researcher to compare directly the treatment versus control effect sizes across conditions of another variable. In this way the researcher could more closely test the hypothesis of interest with a single test while other post hoc procedures may address the hypothesis of interest only indirectly through multiple comparisons. While the ANOVA interaction test provides a single significance test that provides similar information to the effect size based test, it

does not directly test the effect sizes themselves.

If the effect size measure known as Hedges' g (Hedges & Olkin, 1985) is used, each of the independent effect sizes would be calculated as follows:

$$g = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{((n_1 - 1)s_1^2) + ((n_2 - 1)s_2^2))}{(n_1 + n_2 - 2)}}} \qquad \textbf{(1)}$$

where $\underline{X}_1$ is the sample mean for group 1, $\underline{X}_2$ is the sample mean for group 2, $\underline{s}^2_1$ is the sample variance for group 1, $\underline{s}^2_2$ is the sample variance for group 2, $\underline{n}_1$ is the sample size for group 1, and $\underline{n}_2$ is the sample size for group 2. The sampling variance of Hedges' g effect size measure takes the form (Rosenthal, 1994):

$$\sigma_g^2 = \frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2(n_1 + n_2 - 2)}. \qquad \textbf{(2)}$$

Therefore, a test for the difference between independent effect sizes using this method could take the form:

$$z = \frac{g_1 - g_2}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{g_1^2}{2(n_1 + n_2 - 2)} + \frac{n_3 + n_4}{n_3 n_4} + \frac{g_2^2}{2(n_3 + n_4 - 2)}}} \qquad \textbf{(3)}$$

where g_1 represents the effect size for groups 1 and 2, g_2 represents the effect size for groups 3 and 4, n_1 represents the sample size for group 1, n_2 represents the sample size for group 2, n_3 represents the sample size for group 3, and n_4 represents the sample size for group 4.

If we refer to the population effect size as δ, g has been shown to be a biased estimator of δ (Hedges, 1981, Hedges & Olkin, 1985). Hedges (1981) suggested using d in place of g, which is obtained from a procedure that approximates a correction for this bias:

$$d \cong \left(1 - \frac{3}{4N - 9}\right) g \qquad (4)$$

where $N = n_1 + n_2$.

The sampling distribution of the standardized mean difference effect size measure d has been shown to be non-central t (Hedges 1981, 1982). Hedges and Olkin (1985) offer the sampling variance of d:

$$\sigma_d^2 = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}. \qquad (5)$$

Simulation efforts suggest that if the sampling distribution of d is assumed to be normal and sample size is at least moderate, the resulting confidence limits would be very similar to those obtained when using the appropriate non-central t distribution (Hedges, 1981, 1982; Hedges & Olkin, 1985). Researchers who are interested in the exact method as proposed by Hedges and Olkin (1985) should see Steiger and Fouladi (1997) for a very useful discussion that explains

how to use the appropriate non-central t distribution to form confidence intervals around and make comparisons between effect sizes. This paper addresses approximate methods that are less computationally intensive.

Rosenthal and Rubin (1982) outline a method for testing the differences among a series of effect size estimates by utilizing linear contrasts and a test statistic distributed as approximately normal. They offer an analytical solution that specifies the nature of the differences between the normal and more precise non-central t methods. If this method is collapsed to a pairwise comparison, it is very similar to a z test format. Alliger (1995) tested an application of the z test method to the difference between two effect size estimates obtained by summarizing a series of primary studies. The test performed well, showing both Type I and Type II error rates to be very similar to what would be expected if the test statistic were actually distributed as normal. Results such as these suggest that a z statistic for the difference between standardized mean difference effect sizes could be treated as if it were normally distributed (Gleser & Olkin, 1994).

A test for the difference between independent effect sizes would be calculated as follows:

$$z = \frac{d_1 - d_2}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d_1^2}{2(n_1 + n_2)} + \frac{n_3 + n_4}{n_3 n_4} + \frac{d_2^2}{2(n_3 + n_4)}}} \qquad (6)$$

where d_1 represents the effect size for groups 1 and 2 and d_2 represents the effect size for groups 3 and 4. This procedure is equivalent to a special case of the homogeneity of effect size test proposed by Hedges and Olkin (1985) for the case when there are only two effect sizes.

Dependent Group Means

The dependent case arises if we allow the means involved in an effect size calculation to be dependent, while assuming that the effect sizes themselves for two different groups remain independent. This is the situation faced when using the Split Plot Repeated Measures ANOVA design with two groups (e.g., treatment and control), and two occasions (e.g., pre and post-tests). There are a variety of effect size calculation methods for dependent group means (Tong & Shadish, 1996). However, many of them involve the use of individual gain scores, requiring a change in the scaling of the effect size estimates from units of the standard deviation of the dependent variable to some other metric such as the standard deviation of gain scores. When the scaling of the effect size metric is no longer the standard deviation of the original scores, interpretation is difficult (Becker, 1988). In addition, gain score methods can often have the effect of increasing the magnitude of the effect size estimate as well as making combination with standard effect size estimates problematic. The effect size measure which offers effect size estimates that are scaled in a manner similar to those commonly used with independent means, takes the form (Glass, McGaw & Smith, 1981):

$$g = \frac{\overline{X_{T1}} - \overline{X_{T2}}}{s_{T1}} \qquad (7)$$

where $\underline{X}_{T1}$ represents the sample mean for time 1, $\underline{X}_{T2}$ represents the sample mean for time 2, and s_{T1} represents the sample standard deviation for time 1. The variance of g can be estimated using the formula (Becker, 1988):

$$\sigma_g^2 = \frac{2(1-r)}{n} + \frac{g^2}{2(n-1)} \qquad (8)$$

where r represents the sample correlation between the scores on the dependent variable obtained at time 1 and time 2. Therefore, a dependent z test for g could take the form:

$$z = \frac{g_1 - g_2}{\sqrt{\frac{2(1-r_1)}{n_1} + \frac{g_1^2}{2(n_1-1)} + \frac{2(1-r_2)}{n_2} + \frac{g_2^2}{2(n_2-1)}}} \qquad (9)$$

where r_1 represents the correlation between the scores on the dependent variable obtained at time 1 and time 2 for group 1 and r_2 represents the correlation between the scores on the dependent variable obtained at time 1 and time 2 for group 2.

An unbiased effect size metric for the dependent case can be obtained as follows (Hedges, 1981):

$$d \cong \left(1 - \frac{3}{4n-5}\right) g \qquad (10)$$

where n represents sample size. If this bias correction procedure is applied, the sampling variance of d becomes (Becker, 1988):

$$\sigma_d^2 = \frac{2(1-r)}{n} + \frac{d^2}{2n}. \qquad (11)$$

Therefore, a test for the difference between two dependent effect sizes using the d method could

$$z = \frac{d_1 - d_2}{\sqrt{\frac{2(1-r_1)}{n_1} + \frac{d_1^2}{2n_1} + \frac{2(1-r_2)}{n_2} + \frac{d_2^2}{2n_2}}}. \qquad (12)$$

take the form:

This procedure is also equivalent to a special case of a homogeneity of effect size test proposed by Hedges and Olkin (1985) for the case when there are only two effect sizes.

This study evaluates the Type I error rates and power of both the independent and dependent z tests, using critical values for the test statistic acting as if the test statistics were actually distributed as normal under the condition that the null hypothesis is true. In addition, test statistics based on the g and d methods will be compared.

Method

Monte Carlo simulation was used to generate primary study data from which Type I error rates and power values were obtained for the g and d based z tests. The simulation design for the Type I error rate evaluation completely crossed six sample size conditions (5, 10, 20, 30, 50, and 100) by five δ conditions (.00, .25, .50, .75, and 1.00). The δ conditions represent the case in which the null hypothesis is true, $\delta_1 = \delta_2$, while varying the magnitude of δ_1 and δ_2 simultaneously. The power evaluation completely crossed the same six sample size conditions by four conditions representing the magnitude of the difference between treatment, δ_1, and

control, $\underline{\delta}_2$, conditions (.25, .50, .75, and 1.00). This was accomplished by setting the value of $\underline{\delta}_1$ to 0 while the value of $\underline{\delta}_2$ was varied across the four conditions.

In each cell of the simulation designs, primary study data were generated for 10,000 replications. These data were generated as if an experiment using a 2x2 factorial ANOVA design had been conducted. In the independent case, the design had four cells formed by completely crossing two between-subjects factors. In the dependent case, the design had four cells formed by completely crossing one between-subjects factor and one within-subjects factor. Means, standard deviations, effect sizes, and test statistics were calculated from the primary data. The sample size parameters in the simulation study refer to the cells in these simulated ANOVA designs.

The RANNOR utility within SAS was used to generate samples from normally distributed populations. The sample data for each cell in both designs was generated from population conditions with equal cell variances. In the cases where dependent means effect sizes were calculated, the correlation between the levels of the within-subjects factor was kept constant in the population case by fixing this value at a moderate level, ρ=.5.

Results

Table 1 illustrates that the empirically generated Type I error rates for the d based independent z statistic were found to be consistently closer to a nominal alpha of .05 than those of the g based statistic. The d based statistic yielded Type I error rates that were more conservative than nominal alpha for all but the cells that include sample sizes of 100. However, the d based statistic yielded Type I error rates for the two-tailed test cells of sample size 100 that exceeded the nominal alpha of .05 by no more than .0009. The g based statistic yielded Type I

error rates that exceeded a nominal Type I error rate of .05 in 76.67% of the cells. In all 60 of the cells simulated, the d based statistic yielded lower or more conservative Type I error rates than did the g based statistic. As the magnitude of the difference between population effect sizes increased, the Type I error rates tended to decrease.

Insert Tables 1 and 2 About Here

For the dependent z statistics, Table 2 shows that the d based statistic again yielded more conservative Type I error rates than did the g based statistic in every one of the 60 cells simulated. The d based method yielded Type I error rates that exceeded nominal alpha in the n=5 conditions while remaining more conservative than nominal alpha for almost all the larger sample size cells. The g based method yielded Type I error rates that exceeded nominal alpha in all but three cells. For both the d and g based statistics, as sample size increased, the Type I error rates decreased.

Table 3 reports the empirically generated power values for the independent case. Table 4 reports the empirically generated power values for the dependent case. The g method yielded power values that were greater than or equal to those for the d method in every cell of the design for the both the cases of independent and dependent means. The gap between the g and d methods lessons as sample increases. These results must be taken in the context of higher Type I error rates for the g method. Since the actual alpha level for the g method is both higher than the d method and higher than nominal alpha, some difference in power favoring the g method would be expected. As expected, the power of the one-tailed tests is greater than the two-tailed tests for

every condition in this study.

Insert Tables 3 and 4 About Here

Power of .80 is considered adequate for experiments in educational research (Cohen, 1988). The case of independent means only approaches this level, as indicated by the bolded values on Table 3, when sample size is 30 and the difference between the population effect sizes is at least 1.0. For sample sizes of 50, this level is approached, as the difference is at least .75 while for sample sizes of 100, while it is approached with differences of .50. The case of dependent means approaches this level more often. As indicated by the bolded values on Table 4, adequate power levels are approached when sample size is 20 and the difference between the population effect sizes is at least .75. For sample sizes of 30, this level is approached when the difference is at least .75. However, for sample sizes of 50 and 100, it is approached with differences of .50.

Discussion

The results of this investigation showed that the d based statistic yielded Type I error rates that were closer to the nominal level than did the g based statistic. The d based statistic appears to be a slightly conservative method for testing the difference between two independent standardized mean difference effect size measures. The d based statistic tests the difference between effect size indexes that in a sense have already been adjusted for sample size through the bias correction factor. Thus, Type I error rates in the independent case were very close to nominal even for small sample size situations. For the dependent case, the d based statistic was

unable to remain close to the nominal alpha level at sample sizes of 5. However, the g based statistic was unable to remain close to nominal alpha in any of the situations tested, particularly the small sample size conditions where empirical alpha approached .10 in the worst case.

It should be noted that while the dependent effect size measure evaluated has the advantage of removing the effects of history, retesting, and maturation by subtracting from the treatment group effect that of the control, it suffers from the weakness of the potentially unrealistic assumption of independence of effect sizes. It may be unrealistic given that a well executed design would in fact strive to obtain equivalence between treatment and control groups which can lead to similarity and therefore covariance between effect sizes when viewed across a sample of effect sizes from various studies (Becker, 1988).

The power evaluation for the independent case d based statistic show that power is quite low in situations with sample cell sizes of less than 30. In addition, the population difference in effect sizes had to be at least .50, even in the largest sample size conditions tested for power to reach acceptable levels. For the dependent d based statistic, power was quite low for sample sizes less than 20 and again the population difference in effect sizes had to be at least .50, even in the largest sample size conditions tested for power to reach acceptable levels. These results suggest that this procedure will have some limitations as posthoc comparison technique when factorial ANOVA designs have small cell sizes.

This simulation evaluated power by examining situations in which the control group had a population effect size of zero. Future research could extend this work to look at the performance of these statistics when there is a population difference between treatment and control group effect sizes and the control group population effect size is non-zero. In addition,

only primary study data conditions that included normality, equal sample sizes, and equal variances were considered. Future research could focus on the robustness of these procedures to primary data conditions that include non-normality and inequalities of sample size and variance. For the case of effect sizes calculated based on dependent means, future research could focus on varying the degree of correlation between observations across levels of the within-subjects term. While this study focused on only 2x2 ANOVA designs, future efforts could also focus on designs that yield more than two effect sizes. Application of this procedure to such designs would benefit from an extension of these methods that allows for multiple comparisons while avoiding the inflated Type I error rate problem. Future research could focus on evaluating an adaptation of the methods proposed by Hedges and Olkin (1985) that involve an extension of Scheffe's post hoc comparison approach (1953, 1959) to the comparison of effect sizes.

The results of this simulation offer evidence of the usefulness of the z statistics presented for comparing independent and dependent d based effect size metrics. Considering the reasonable Type I error rates and power levels for the d based statistic when sample sizes are at least moderate (n/cell=20-30), these procedures present some advantages to educational researchers who use 2x2 factorial designs. They require the researcher to calculate effect sizes, to recognize the influence of sampling error on effect sizes, and to think about interactions in terms of the differences between pairs of standardized mean differences. Furthermore, they present the possibility of running only a single test to enhance interpretation of interactions.

Applications to Educational Research

Example of Independent Group Means Case

Data from a study of Head Start children (Kalabaca, Lambert, Abbott-Shim, & Springs,

2001) were used to illustrate the independent group means case. The researchers were interested in examining whether the father's presence or absence in the home had a differential effect on prosocial behavior across children who had been exposed to home violence and children not exposed to home violence. A 2X2 ANOVA was calculated, where the dependent variable was prosocial behavior as reported by the child's teacher and the independent variables were father's presences in the home (yes or no) and child's exposure to home violence or criminality (yes or no). The means, standard deviations, and sample sizes for prosocial behavior for father presence across child's expose to violence are reported in Table 5. There was a statistically significant main effect for father presence (F=11.21, p<.01) and a statistically significant interaction (F=7.17, p<.01). There was not a statistically significant main effect for home violence (F=0.92, p>.05). Tukey post hoc procedure indicated that children had higher prosocial scores when the father was present in the nonviolent homes and in the violent homes when compared to children in the father absent in the violent home condition.

Using an effect size method, the researchers could calculate the d for the no home violence condition for the difference between the father present and father absent conditions (d=0.09) and similarly for the same difference in the home violence condition (d=0.80). To examine differential effects for presence or absences of father, the researcher could test for the differences between the two effect sizes (z=-2.70, p≤.01). This single test would illustrate that the father absence is associated with fewer prosocial behaviors for children living in violent homes than it is for children living in non-violent homes.

Insert Tables 5 and 6 About Here

Example of Dependent Group Means Case

Wilkes, Lambert, and Vanderwillie (1998) investigated the effect of providing technical assistance to family daycare providers following inspection of their facilities. Half of a random sample of providers from the state of Georgia were randomly assigned to receive the assistance while the remaining sites received no assistance. The pretest and posttest scores represent the percent correct scores on an observational measure of their compliance with state regulations for family daycare providers. Table 6 displays the means and standard deviations for each observation. The central research question involved examining whether the group receiving the technical assistance treatment would make greater gains in compliance from pretest to posttest than the control group. The main effect for group was not statistically significant ($F=.63, p>.05$). There was a statistically significant main effect for time ($F=539.46, p<.001$) and a statistically significant interaction between group and time ($F=64.19, p<.001$). Tukey post hoc comparison procedures indicated that both groups had posttest means greater than their pretest means and the groups were neither equivalent at pretest or posttest.

Using an effect size method, the researchers could calculate the $\underline{d}$ for both the experimental group ($\underline{d}=1.18$) and control group ($\underline{d}=0.59$). To examine differential effect from pretest to posttest for the experimental and control groups, the researchers could test for the differences between the two effect sizes ($\underline{z}=7.42$, $\underline{p}<.01$). This single test would clearly indicate that the effect size for the treatment group was greater than the effect size for the control group,

indicating that technical assistance has a positive effect on increases in compliance with state regulations beyond what would be expected by monitoring compliance without any further assistance to providers.

In both the independent and dependent means cases illustrated here, the single significance test gave a clear indication of the answer to the central research question under investigation. This method also facilitated the use of effect sizes that enhance the interpretability of the results.

References

Alliger, G. M. (1995). The small sample performance of four tests of the difference between pairs of meta-analytically derived effect sizes. Journal of Management, 21, 789-799.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

American Psychological Association. (2001). Publication manual of the American Psychological Association (5th ed.). Washington, DC: Author.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology, 41, 257-278.

Becker, B. J. (1989, March). Model-driven meta-analysis: possibilities and limitations. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Becker, B. J. & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges, The handbook of research synthesis. New York: Russell Sage Foundation.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Erlbaum.

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & Psychology, 5, 75-98.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. The Journal of Educational Research, 94(5), 275-282.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a

series of experiments involving paired comparisons. Journal of Educational Statistics, 18, 271-279.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly: Sage.

Gleser, L. J. & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges, The handbook of research synthesis. New York: Russell Sage Foundation.

Harlow, L., Mulaik, S., & Steiger, J. (eds.) (1997). What if there were no significance tests? New York: Erlbaum.

Hedges, L. V. (1981). Distribution theory for Glass=s estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490-499.

Hedges, L. V. & Olkin, I. (1985). Statistical methods for meta-analysis. San Diego, CA: Academic Press.

Kalabaca, I., Lambert, R., Abbott-Shim, M., & Springs, J. (2001). Father presence, exposure to violence, and the social functioning of Head Start children. Paper accepted for presentation to the Head Start National Research Conference, Washington, D.C.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Levin, J. (1993). Statistical significance testing from three perspectives. The Journal of Experimental Education, 61, 378-382.

Miller, N. & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H.

Cooper & L. V. Hedges, The handbook of research synthesis. New York: Russell Sage Foundation.

Oshima, T., & McCarty, F. (2000, April). Factorial analysis of variance statistically significant interactions: What's the next step? Paper presentation, American Educational Research Association Annual Meeting, New Orleans, LA.

Rosenthal, R. & Rubin, D. B. (1982). Comparing effect sizes of independent studies. Psychological Bulletin, 92, 500-504.

Scheffe, H. (1953). A method for judging all contrasts in the analysis of variance. Biometrika, 40, 87-104.

Scheffe, H. (1959). The analysis of variance. New York: John Wiley & Sons.

Thompson, B. (1993). The use of statistical significance in research: Bootstrap and other alternatives. The Journal of Experimental Education, 61, 361-377.

Thompson, B. (1996). AERA Editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26-30.

Tong, Z. & Shadish, W. R. (1996, August). The comparison between exact and inexact effect sizes computed by change score methods. Paper presented at the meeting of the American Psychological Association, Toronto, Ont.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? American Psychologist, 24, 83-91.

Wilkes, D., Lambert, R., & Vanderwillie, L. (1998). Technical assistance as part of routine inspections of family child care homes. Early Childhood Research Quarterly, 13(2), 355-372.

Table 1.
Empirically Generated Type I Error Rates for the Independent z Test

Test Statistic	δ_1, δ_2,	Sample Size Per Cell					
		5	10	20	30	50	100
One Tailed Test - z_d	0.00	.0465	.0448	.0465	.0490	.0489	.0533
	0.25	.0462	.0453	.0464	.0488	.0495	.0537
	0.50	.0461	.0467	.0456	.0488	.0486	.0525
	0.75	.0464	.0471	.0466	.0493	.0489	.0529
	1.00	.0460	.0481	.0468	.0489	.0484	.0522
Two Tailed Test - z_d	0.00	.0477	.0476	.0477	.0487	.0487	.0501
	0.25	.0459	.0468	.0472	.0480	.0487	.0508
	0.50	.0466	.0463	.0463	.0469	.0480	.0509
	0.75	.0474	.0470	.0464	.0482	.0477	.0508
	1.00	.0479	.0474	.0468	.0475	.0481	.0496
One Tailed Test - z_g	0.00	.0607	.0516	.0503	.0518	.0496	.0540
	0.25	.0592	.0524	.0490	.0511	.0509	.0545
	0.50	.0583	.0533	.0487	.0505	.0495	.0535
	0.75	.0559	.0524	.0486	.0509	.0503	.0534
	1.00	.0546	.0526	.0483	.0506	.0496	.0528
Two Tailed Test - z_g	0.00	.0657	.0549	.0515	.0509	.0497	.0508
	0.25	.0644	.0560	.0507	.0506	.0503	.0514
	0.50	.0616	.0546	.0506	.0495	.0499	.0513
	0.75	.0597	.0552	.0490	.0508	.0500	.0510
	1.00	.0577	.0541	.0498	.0488	.0488	.0504

Table 2.
Empirically Generated Type I Error Rates for the Dependent z Test.

Test Statistic	$\delta_1, \delta_2,$	Sample Size Per Cell					
		5	10	20	30	50	100
One Tailed Test - z_d	0.00	.0554	.0481	.0463	.0491	.0471	.0519
	0.25	.0531	.0489	.0470	.0481	.0475	.0518
	0.50	.0525	.0491	.0460	.0489	.0508	.0510
	0.75	.0526	.0497	.0472	.0491	.0505	.0509
	1.00	.0525	.0511	.0494	.0496	.0504	.0508
Two Tailed Test - z_d	0.00	.0601	.0520	.0492	.0508	.0476	.0510
	0.25	.0618	.0494	.0471	.0503	.0469	.0499
	0.50	.0597	.0490	.0506	.0483	.0484	.0499
	0.75	.0562	.0483	.0493	.0483	.0479	.0505
	1.00	.0523	.0490	.0476	.0494	.0477	.0472
One Tailed Test - z_g	0.00	.0784	.0599	.0528	.0538	.0500	.0530
	0.25	.0783	.0600	.0517	.0529	.0502	.0527
	0.50	.0711	.0584	.0504	.0534	.0524	.0522
	0.75	.0640	.0571	.0512	.0514	.0524	.0513
	1.00	.0591	.0548	.0514	.0512	.0512	.0515
Two Tailed Test - z_g	0.00	.0944	.0680	.0578	.0557	.0511	.0527
	0.25	.0896	.0670	.0557	.0542	.0490	.0514
	0.50	.0795	.0634	.0566	.0528	.0510	.0512
	0.75	.0659	.0580	.0531	.0515	.0502	.0509
	1.00	.0545	.0547	.0501	.0516	.0493	.0483

Table 3.
Empirically Generated Power Values for the Independent z Test.

Test Statistic	$\delta_1 - \delta_2$	Sample Size Per Cell					
		5	10	20	30	50	100
One Tailed Test - z_d	0.25	.0771	.0979	.1315	.1597	.2202	.3402
	0.50	.1248	.1811	.2839	.3821	.5426	**.7985**
	0.75	.1865	.2987	.4944	.6482	**.8302**	**.9803**
	1.00	.2632	.4325	.6985	**.8498**	**.9647**	**.9993**
Two Tailed Test - z_d	0.25	.0568	.0608	.0798	.1028	.1435	.2337
	0.50	.0774	.1139	.1867	.2632	.4141	**.7023**
	0.75	.1132	.1979	.3655	.5233	**.7406**	**.9592**
	1.00	.1671	.3116	.5781	**.7625**	**.9335**	**.9981**
One Tailed Test - z_g	0.25	.1005	.1113	.1390	.1647	.2241	.3426
	0.50	.1526	.1994	.2978	.3898	.5484	**.8013**
	0.75	.2244	.3186	.5061	.6556	**.8334**	**.9804**
	1.00	.3034	.4560	**.7094**	**.8553**	**.9655**	**.9993**
Two Tailed Test - z_g	0.25	.0734	.0726	.0861	.1065	.1462	.2362
	0.50	.1003	.1274	.1962	.2713	.4188	**.7044**
	0.75	.1459	.2186	.3793	.5336	**.7456**	**.9599**
	1.00	.2040	.3371	.5916	**.7700**	**.9353**	**.9981**

Table 4.
Empirically Generated Power Values for the Dependent z Test.

Test Statistic	δ_1 δ_2	Sample Size Per Cell					
		5	10	20	30	50	100
One Tailed Test - z_d	0.25	.0999	.1276	.1853	.2369	.3403	.5416
	0.50	.1717	.2659	.4460	.5890	**.7838**	**.9664**
	0.75	.2596	.4447	**.7205**	**.8694**	**.9742**	**.9997**
	1.00	.3671	.6244	**.8940**	**.9744**	**.9989**	**1.0000**
Two Tailed Test - z_d	0.25	.0705	.0785	.1139	.1503	.2312	.4141
	0.50	.1105	.1703	.3196	.4608	.6778	**.9305**
	0.75	.1708	.3170	.5995	**.7834**	**.9476**	**.9988**
	1.00	.2489	.4902	**.8195**	**.9457**	**.9965**	**1.0000**
One Tailed Test - z_g	0.25	.1454	.1539	.2012	.2491	.3485	.5474
	0.50	.2326	.3076	.4692	.6065	**.7900**	**.9668**
	0.75	.3393	.4900	**.7399**	**.8773**	**.9751**	**.9997**
	1.00	.4466	.6658	**.9018**	**.9764**	**.9990**	**1.0000**
Two Tailed Test - z_g	0.25	.1116	.1017	.1294	.1618	.2401	.4190
	0.50	.1598	.2060	.3421	.4799	.6864	**.9325**
	0.75	.2321	.3679	.6271	**.7951**	**.9494**	**.9989**
	1.00	.3220	.5404	**.8345**	**.9495**	**.9967**	**1.0000**

Table 5.
Prosocial Behavior of Preschool Children By Father Presence and Exposure to Violence.

		Father Present	Father Absent	Effect Size	95% Lower Limit	95% Upper Limit
Not Exposed to Violence	Mean	57.58	56.86	0.09	-0.24	0.41
	SD	7.83	10.19			
	n	158	47			
Exposed to Violence	Mean	59.43	52.94	0.80	0.40	1.21
	SD	7.60	9.02			
	n	82	36			

Note. Results of z Test of the Difference Between Effect Sizes: z=2.698, p=.007.

Table 6.
Percent Compliance For Experimental and Control Groups.

		Pretest	Posttest	Effect Size	95% Lower Limit	95% Upper Limit
Experimental Group	Mean	82.36	93.71	1.18	1.05	1.32
	SD	9.58	7.41			
	n	362	362			
Control Group	Mean	85.70	91.23	0.80	0.40	1.21
	SD	9.39	9.04			
	n	362	362			

Note. Results of z Test of the Difference Between Effect Sizes: z=6.697, p=.000.